RUNNING HEAD: The (Dis-) Confirmability of Warmth and Competence

The Confirmability and Disconfirmability of Trait Concepts Revisited: Does Content Matter?

Nicole Tausch

University of Oxford, UK

Jared Kenworthy

University of Texas at Arlington

and

Miles Hewstone

University of Oxford, UK

Citation: Tausch, N., Kenworthy, J., & Hewstone, M. (2007). The confirmability and disconfirmability of trait concepts revisited: Does content matter? *Journal of Personality and Social Psychology, 92*, 554-556.

Abstract

Rothbart and Park (1986) demonstrated that, consistent with the common negativity bias, positive traits are difficult to confirm and easy to disconfirm while the opposite is true for negative traits. We extend their analysis by showing that trait (dis-)confirmability is moderated by trait content (Warmth vs. Competence). Study 1 identifies a trait sample representative of Warmth and Competence. Study 2 shows a strong negativity effect for Warmth, and a reduced (or absent) negativity effect for Competence. Study 3 examines trait properties related to the behavioural range of the trait-possessor and to the motivational goals of the perceiver as predictors of trait (dis-) confirmability. The theoretical and practical implications of our findings are discussed and avenues for future research are suggested.

KEY WORDS: social perception, trait attribution, trait disconfirmability, behaviour diagnosticity, motivation

Judging other people's personality traits is a common task people face in everyday life. Trait ascriptions help us to explain others' behaviour, predict their future behaviour, and guide our own behaviour towards them. The rules by which people infer personality traits on the basis of observed behavioural instances have long been of interest to social psychologists (see Gilbert, 1998, for a review). Many authors have stressed the role of behaviour diagnosticity in this process. For example, Jones and Davis (1965) suggested in their classic correspondent inference theory that potentially costly behaviour that is at odds with situational demands or that violates social norms is particularly informative about some underlying stable quality of the actor and will thus lead to confident dispositional attributions. Socially desirable behaviour on the other hand is likely to be performed no matter whether an actor possesses the corresponding disposition or not (because it is socially rewarded), and is thus not very informative about the actor's true disposition.

While most work on trait attribution has focused on the inferential principles that guide causal reasoning, a number of studies have investigated the attributes *inherent* in trait concepts as determinants of how traits are diagnosed (e.g., Funder & Dobroth, 1987; Reeder & Brewer, 1979; Rothbart & Park, 1986). Reeder and Brewer (1979) proposed that the trait inference process may be affected by variations in the schematic representation of traits, in particular the implicational links between dispositional levels and relevant behaviours. They proposed that each position of a target person on a bipolar dispositional continuum is implicationally associated with a range of behaviours on the corresponding behavioural attribute continuum. Reeder and Brewer outlined three schemata of association (partially restrictive, hierarchically restrictive, and fully restrictive), each having different implications for the rules of inference used when making trait attributions based on observed behaviours.

Perhaps most interesting is the hierarchically restrictive schema, which implies an asymmetrical association of dispositional levels and possible behaviours: Individuals at one

dispositional extreme are associated with a wider range of behaviours than are individuals at the other extreme. According to Reeder and Brewer (1979), this schema is likely to apply to ability attributes as well as moral attributes. For example, *unintelligent* individuals are expected to be capable of only unintelligent behaviour, whereas the behavioural range of *intelligent* people may be greater due to motivation and task demands and can therefore include both intelligent and unintelligent behaviour. Consequently, a single intelligent behaviour can lead to a confident attribution of the trait *intelligent* to an actor, but a single unintelligent behaviour is less diagnostic about the actor's disposition (see Reeder, 1979, for empirical evidence). The hierarchically restrictive schema can be similarly applied to moral traits. For example, *honest* individuals are expected to engage almost exclusively in honest behaviours, whereas *dishonest* individuals could engage in both honest and dishonest behaviours. A single dishonest behaviour can thus be sufficient to lead us to attribute the trait *dishonest* to an actor, whereas a single honest behaviour is less informative (Reeder & Spores, 1983).

Based in part on these ideas, Rothbart and Park (1986) investigated the 'confirmability' (the ease with which a trait is instantiated) and 'disconfirmability' (the ease with which a trait ascription is revised) of trait concepts and demonstrated that different traits have different confirmation and disconfirmation thresholds. They showed that, as a general rule, 'bad' reputations are easy to gain and difficult to lose, whereas the opposite is true for 'good' reputations. In this paper we revisit the issue of trait (dis-) confirmability. We argue that Rothbart and Park's analysis can be further extended and refined by a systematic investigation of trait content, specifically by distinguishing between the two core dimensions of social perception, Warmth and Competence (e.g., Fiske, Cuddy, Glick, & Xu, 2002; Wojciszke, Bazinska, & Jaworski, 1998a). Before outlining the present studies and our

predictions, we briefly review Rothbart and Park's classic study and the relevant literature on Warmth and Competence as central dimensions of social perception.

The Confirmability and Disconfirmability of Trait Concepts

Rothbart and Park (1986) argued that personality traits differ on at least three relevant dimensions which may determine their diagnosis. First, Rothbart and Park suggested that the evidence-to-inference link is not identical for all traits (i.e., for some traits, a corresponding behaviour is more informative than for others). This dimension relates to the diagnosticity aspect of behavioural information discussed above (Jones & Davis, 1965; Reeder & Brewer, 1979) and was operationalized by assessing the number of behavioural instances required to confirm a trait and the number of behavioural instances required to disconfirm it.

Second, Rothbart and Park (1986) proposed that trait concepts vary in the degree to which they imply clear and specific behavioural referents. For example, it seems that confirmatory and disconfirmatory behaviours can be more clearly specified for traits such as *friendly* than for traits like *sly*. According to Rothbart and Park, traits may vary on this dimension due to their abstractness, their level of generality, or simply due to the number of clear behavioural exemplars accessible in memory. Third, Rothbart and Park (1986) suggested that the frequency with which occasions arise in everyday life that allow for confirming or disconfirming behaviour may vary between different types of traits. This dimension relates to the structural aspects of the social environment that determine the occurrence of different kinds of behaviour and thus the likelihood with which trait attributions can be made. For example, there are probably many more social occasions that allow for *friendly* behaviours than occasions that allow for *heroic* behaviours. Correspondingly, there are fewer occasions to engage in behaviour that disconfirms the trait *heroic* than occasions to engage in behaviour that disconfirms the trait *friendly*.

Treating traits as the unit of analysis, Rothbart and Park (1986) asked independent groups of judges to rate 150 personality traits on (1) the number of behavioural instances required to confirm a trait and the number of instances required to disconfirm it (*Instances Confirming* and *Instances Disconfirming*), (2) the ease of imagining confirming and disconfirming behaviours (*Imaginability Confirming* and *Imaginability Disconfirming*), and (3) the frequency of occasions that allow for the confirmation and disconfirmation of a trait (*Occasions Confirming* and *Occasions Disconfirming*). In addition, the traits were scaled according to their favourability as well as their perceived frequency in the general population.

Rothbart and Park (1986) found an interesting pattern of intercorrelations between the different dimensions. They obtained positive correlations between the favourability of a trait and its judged population frequency, as well as the frequency of occasions that allow for confirming behaviours, reflecting the high frequency and thus low information value of positive events (Jones & Davis, 1965; Kanouse & Hanson, 1972). There were also positive correlations between the occasions and imaginability dimensions, suggesting that frequently occurring traits tend to have clearer behavioural referents. Furthermore, while the occasions and imaginability dimensions showed symmetrical relationships (i.e., traits for which confirming behaviours can occur frequently/were easy to imagine also tended to be the traits for which disconfirming behaviours can occur frequently/were easy to imagine), a clear asymmetry was found for the instances dimensions. The negative correlation between the instances confirming and instances disconfirming ratings (r = -.71) suggests that traits which are easy to gain are harder to lose and vice versa.

Moreover, whereas the instances confirming dimension correlated positively with favourability (r = .71), instances disconfirming correlated negatively with favourability (r = -.70). This implies that unfavourable traits are easy to acquire and hard to lose, and that favourable traits are hard to acquire and easy to lose. This finding is consistent with the

common negativity bias, which is the greater impact of evaluatively negative than of equally intense positive stimuli (see Baumeister, Bratslavsky, Finkenhauer, & Vohs, 2001, for a review). This bias has been explained in a number of ways. While some accounts expect negativity biases due to the asymmetrical diagnosticity of positive and negative behaviours (e.g., Jones & Davis, 1965), others (range-frequency explanations, see Kanouse & Hanson, 1972) have stressed that negative events impact more heavily on impressions because they are in greater contrast when compared to the slightly positive (but psychologically neutral) point of expectation, and thus appear more extreme and novel than positive events.

Studies that have disentangled frequency from negativity (e.g., Pratto & John, 1991) suggest, however, that there may be an automatic vigilance mechanism which monitors potentially threatening information, resulting in biased attention to and memory for negative information. Peeters (1983) gave a functional account of this mechanism in his 'behavioural-adaptive theory'. He suggested that evolutionary pressures underlie both the common general positive orientation (i.e., the Pollyanna principle; Matlin & Stang, 1978) and the negativity bias. The general tendency to expect positive outcomes results in the approaching of novel objects and situations and is functional as it expands the range of beneficial dealings with the environment. However, as uncontrolled approach behaviour would sooner or later have harmful effects, it is coupled with a strong sensitivity to aversive stimuli and overemphasis of the negative. This is connected to a straightforward avoidance reaction, which serves to avert danger quickly (see Peeters & Czapinski, 1990, for a review of evidence).

The Content of Social Perception: Warmth and Competence

The social perception literature suggests that when people interact with others they are mainly interested in finding out (a) what others' goals are in relation to their goals (i.e., others' *intent* toward them, e.g., whether they are *trustworthy*, *tolerant*, *deceitful*, or *aggressive*) and (b) whether they are capable of realizing those goals (e.g., whether they are

intelligent, *ambitious*, *stupid*, or *lazy*). Fiske et al. (2002) have labelled these dimensions Warmth and Competence, respectively.

The two dimensions were first established in research on implicit personality theories which, using multidimensional scaling, obtained the dimension of *social desirability*, which encompassed traits like *honest*, *helpful*, and *sincere* at one end and *dishonest*, *selfish* and *irresponsible* at the other, and the dimension *intellectual desirability*, which contained traits such as *intelligent*, *persistent*, and *skilful* at one end, and *foolish*, *unintelligent*, and *clumsy* at the other (Rosenberg, Nelson, & Vivekanathan, 1968). Similarly, Peeters (1983) proposed that traits group along two dimensions: that of *self-profitability,* which pertains to Competence and is related to the adaptive value of attributes for the self, and *other-profitability*, which pertains to Warmth and is related to the adaptive value of attributes for others. He thus distinguished four groups of traits: (1) positive self-profitable traits, which have maximum positive outcomes for the trait-possessor, such as *confident* or *intelligent*; (2) negative self-profitable traits, which are generally disadvantageous for the trait-possessor, for example, *slow* or *unintelligent*; (3) positive other-profitable traits, which have a beneficial effect on other people surrounding the trait possessor, such as *trustworthy* and *tolerant*; and (4) negative other-profitable traits, which have a harmful effect on other people surrounding the trait possessor, for example *selfish* and *intolerant*.

Wojciszke and colleagues (see Wojciszke, 2005, for a review) demonstrated in a series of studies that Warmth and Competence are two basic kinds of content used when construing actions, evaluating others, and perceiving the self. Furthermore, consistent with Peeters' proposition, Wojciszke, Dowhyluk, and Jaworski (1998b) showed that Warmth-related traits are generally perceived as being more other- than self-profitable and that Competence-related traits are perceived as being more self- than other-profitable. Warmth and Competence are also central components of group stereotypes. For example, Phalet and

Poppe (1997) found that Warmth (*honest*, *tolerant*, *modest*, *aggressive*, *selfish*, and *rude*) and Competence (*efficient*, *competitive*, *self-confident*, *intelligent*, *slow*, and *clumsy*) emerged as distinct dimensions of group stereotypes (see also Fiske et al., 2002).

A number of accounts specifically predict differences between Warmth and Competence in terms of the rules by which traits are ascribed. Based on Peeters' (1983) behavioural-adaptive approach, Peeters and Czapinski (1990) proposed that, since other people's self-profitable (i.e., Competence-related) qualities are only consequential to the extent that they lead to higher or lower efficiency in inflicting harm or furnishing benefits, only the other-profitable (i.e., Warmth) dimension is unequivocally associated with approach and avoidance. Thus, they proposed that the negativity bias is expected primarily for this dimension and not the self-profitable dimension. The cue-diagnosticity model of impression formation (Skowronski & Carlston, 1987), however, expects opposite patterns of information integration for Warmth and Competence, due to differences in the diagnosticity of positive and negative behaviours for these two dimensions. Skowronski and Carlston proposed that a trait judgement can be compared to a category decision task where behaviours serve as cues and traits as categories. Behaviours that are frequently and almost exclusively associated with members of one category are highly diagnostic of category membership and should lead to confident trait ascriptions. Note that the implications of this model are similar to those of Reeder and Brewer's (1979) schematic model of dispositional attribution.

The cue-diagnosticity model can account for the negativity bias in trait ascription, but also explicitly predicts a *positivity* bias for traits related to ability, for which positive cues may be regarded as more diagnostic than negative ones, as success is generally perceived as evidence that the actor has the ability to perform the task, whereas failure is more ambiguous and may be due to causes other than ability, such as motivational and situational factors. This positivity bias has been reported in a number of studies. For example, Skowronski and

Carlston (1987, 1992) showed that *intelligent* behaviours were seen as more useful when making intelligence judgments than were *unintelligent* behaviours, demonstrating a positivity bias.

The Present Research

We present three studies in this paper. Study 1 was conducted to identify trait adjectives that broadly represent the Warmth and Competence domains. Study 2 examined the obtained trait sample along the original eight dimensions investigated by Rothbart and Park (1986) and aimed to extend their research by testing for the moderating role of trait content on trait confirmability and disconfirmability. Two of Rothbart and Park's dimensions, the number of instances required to confirm a trait and the number of instances required to disconfirm a trait, were of primary importance to this research. An important finding of Rothbart and Park's study was that positive trait ascriptions are hard to gain and easy to lose whereas the opposite is true for negative trait ascriptions. Rothbart and Park did not systematically investigate whether these findings apply equally to different types of traits. They noted, however, that traits relating to ability (e.g., *musical*, *scientifically-minded*, *intelligent, wise*) did not exhibit the general pattern of many-instances-to-confirm/few-instances-to-disconfirm found for positive traits. These traits needed relatively fewer instances to be confirmed and one trait (*musical*) even exhibited the opposite pattern of few-instances-to-confirm/many-instances-to-disconfirm. Nonetheless, Rothbart and Park concluded that "the overwhelming negativity of the majority of traits in this cell seems to argue for the relative importance of favorability as a determinant of dispositional inference" (p. 138). Study 2 sought to clarify whether trait favourability is a sufficient predictor of trait disconfirmability, or whether trait content, specifically Warmth vs. Competence, does matter.

In Study 3, we added five additional dimensions to the analysis to gain further insights into the cognitive and motivational factors underlying the (dis-) confirmability of

Warmth and Competence: (1) the difficulty of pretending to have a trait, (2) the difficulty of hiding a trait, (3) the likelihood of trait-inconsistent behaviour, (4) the potential harmfulness of incorrect trait ascriptions, and (5) the self-desirability of traits. We tested a number of hypotheses concerning differences between Warmth and Competence on these dimensions and assessed the extent to which these factors relate to trait (dis-) confirmability.

Study 1: Identifying a Trait Sample Representative of Warmth and Competence

Positive and negative trait adjectives related to Warmth and Competence were drawn from a pool of items from the person perception and stereotyping literature. To ensure that the trait words chosen for the current research did indeed represent the Warmth and Competence categories, 26 judges were asked to place each trait in either category.

Method

Participants

Participants were 26 (10 male, 14 female, 2 did not specify their gender; mean age = 23 years) students at a British university.

Materials and Procedure

The initial pool of traits included all 150 items from Rothbart and Park's (1986) study and the additional trait words from Anderson's (1968) study on the likeableness of traits. It was further supplemented by traits used in a number of stereotyping studies (e.g., Fiske et al., 2002; Phalet & Poppe, 1997) as well as Rosenberg and Sedlack's (1972) study on implicit personality theories. Adjectives that do not occur in British English (e.g., unentertaining, unagreeable), adjectives that do not represent durable characteristics (e.g., excited, angry), as well as words that are not adjectives (e.g., worrier, liar) were not included. The generated pool of traits contained 641 trait adjectives.

Two independent judges classified these traits into trait words that represent Warmth, Competence, or neither of the two categories. They agreed on 85.9% of these ratings. One

hundred and fifty-three traits were placed by both judges into the Warmth category and 127 traits were consistently placed into the Competence category. These traits were then further rated on their valence (positive vs. negative). The same two judges reached 100% agreement on these ratings. Of the 153 traits related to Warmth, 60 were rated as positive and 93 as negative. Of the 127 Competence-related traits, 71 were rated as positive and 56 as negative.

Next, we examined the frequency of these trait words in British English. The word frequencies were obtained from Kelk (2003) and ranged from 0 for words that are used very infrequently in British English to 16 for words that are used very frequently. Our aim was to choose approximately equal numbers of positive and negative traits in each category, and to have roughly equal average word frequencies in each category. The reason for matching frequency was to rule out the possibility that obtained differences between the content categories on the examined dimensions were due to differential lexical accessibility of the adjectives across the Warmth and Competence dimensions.

First, we deleted all highly uncommon words (word frequency = 0) from the list. Second, highly redundant trait words as well as highly common words which can have additional meanings (e.g., *just*, *cold*) were excluded. The obtained trait list consisted of 30 positive and 33 negative traits related to Warmth, and 30 positive and 32 negative traits related to Competence. Next, the list of 125 trait words was presented to 26 judges, who were asked to sort each trait word into either of the two content categories. The traits appeared in a different random order for each judge. Before making their judgments, judges were provided with a definition of the categories Warmth and Competence. Warmth was defined as representing people's compatibility with other people (whether they are well- or ill-intentioned; beneficial or harmful) and Competence was defined as representing people's ability and capability to complete tasks and achieve status-related goals.

Results and Discussion

The assignment of traits was highly consistent between the 26 judges (mean agreement rate = 91%). For a number of traits, however, the agreement rate was substantially lower. To ensure that traits used in the analyses in Studies 2 and 3 were representative for either Warmth or Competence, we excluded all traits ($N = 31$) for which the agreement rate among judges was below 85%, leaving a sample of 94 traits, 52 of which were related to Warmth and 42 of which were related to Competence. Overall, the final trait list represented the two content categories well and contained the key constructs related to Warmth (e.g., *trustworthy*, *tolerant*, *aggressive*, *selfish*) and Competence (e.g., *intelligent*, *efficient*, *slow*, *incompetent*) used in previous work (e.g., Fiske et al., 2002; Peeters, 1983). A 2 (valence: positive vs. negative) x 2 (content: Warmth vs. Competence) ANOVA on word frequencies yielded no significant effects (all F's <1), confirming that there were no differences in the average word frequency between the categories of trait words in the final trait list.

Study 2: Trait (Dis-) Confirmability and the Moderating Role of Trait Content

Having identified traits representative of the two content categories, we tested the following hypotheses: Consistent with both the behavioural-adaptive perspective (Peeters & Czapinski, 1990) and the diagnosticity approach (Reeder & Brewer, 1979; Skowronski & Carlston, 1987), we expected to replicate Rothbart and Park's finding that positive traits need many instances to be confirmed and few to be disconfirmed, and that negative traits require few instances to be confirmed and many to be disconfirmed for the Warmth domain. The behavioural-adaptive perspective predicts this negativity bias because others' Warmth-related traits have direct consequences for the observer; the behaviour diagnosticity approach expects this negativity bias because of the high diagnosticity of negative behaviours in this domain.

In the Competence domain, however, no such negativity effect is expected. Peeters and Czapinski (1990) proposed that positive-negative asymmetry phenomena involve primarily the other-profitable dimension (i.e., Warmth) and *not* the self-profitable dimension

(i.e., Competence), which is more ambiguously related to approach and avoidance. Thus, these authors predict no negativity bias for Competence ascriptions. However, the diagnosticity approach predicts a positivity effect in the Competence domain because behaviours corresponding to high ability are restricted to people who actually possess that ability (and are thus diagnostic of an actor's disposition), whereas behaviours indicating low ability are not (Reeder & Brewer, 1979; Skowronski & Carlston, 1987). Thus, positive traits should need fewer instances to be confirmed than do negative traits, and negative traits should require fewer instances to be disconfirmed than do positive traits in the Competence domain. No explicit predictions on Warmth-Competence differences on the other dimensions were tested. We assessed these dimensions in order to replicate Rothbart and Park's (1986) results as well as for control purposes. Further extending Rothbart and Park's analysis, this study examined the dimensions' relations to word frequency.

Method

Overview

The procedure for obtaining judgments of trait properties closely followed that of Rothbart and Park (1986). Independent groups of participants rated the trait adjectives obtained in Study 1 on the 8 dimensions assessed in Rothbart and Park's original study. Each participant rated *all* traits (94 target traits and ten practice traits) on *one* dimension only.

Design

Traits, not people, were the unit of analysis. We used a 2 (trait valence: positive vs. negative) x 2 (trait content: Warmth vs. Competence) between-subjects (i.e., traits) design.

Participants

Participants were 81 (27 male, 51 female, 3 did not specify their gender; mean age = 20 years) students at a British university. Each dimension was rated by 9 to 11 participants.

Procedure and Measures

Participants provided their ratings individually. Each participant was randomly assigned to one of the eight judgment tasks – he or she was given the next questionnaire off the top of a randomly mixed stack. Participants were asked to read the detailed instructions printed on each questionnaire very carefully before starting the rating task. The instructions were based as closely as possible on those used by Rothbart and Park (1986). Traits were presented in a different random order for each judge within a given dimension, with the 10 practice items always appearing first. Participants typically completed the task in about 15-20 minutes. Upon completion, they were debriefed and thanked for their participation.

Results and Discussion

Reliability

The first 10 practice items were deleted before the analyses. To determine whether participants were able to make reliable judgments on the dimensions, we followed Rothbart and Park's (1986) practice and calculated the correlation between each participant's ratings and the average ratings by all other participants within that dimension. Participants whose judgments correlated near zero (less than .10) with the average score were excluded. This procedure led to the exclusion of 2 unreliable judges. We then assessed the degree to which the remaining judges agreed, using Cronbach's α. Although this procedure usually examines the degree of interrelatedness among *i* items over *j* judges, we reversed items and judges and assessed the degree of agreement among judges over 94 items. The numbers of original and final judges for each dimension along with coefficient alpha values before and after deletion of unreliable judges appear in Table 1. The final alphas ranged from .66 to .98, with a median of .79, and are similar to those obtained by Rothbart and Park.

Differences between Warmth and Competence on the Dimensions

Using all reliable judges, we computed average scores for each trait on each dimension. Table 2 presents mean ratings for positive and negative Warmth and positive and negative Competence. Results of a series of 2 (valence: positive vs. negative) x 2 (content: Warmth vs. Competence) ANOVAs for each dimension are reported below.

Favourability. The ANOVA yielded a significant main effect of valence, $F(1, 90) = 1345.34$, $p < .001$, such that people possessing positive traits were rated more favourably ($M = 7.15$, $SD = .70$) than people possessing negative traits ($M = 2.99$, $SD = 1.11$), and a main effect of content, $F(1, 90) = 14.44$, $p < .001$, with Competence-related traits rated more favourably ($M = 5.48$, $SD = 1.42$) than Warmth-related traits ($M = 4.86$, $SD = 2.90$). These main effects were qualified by a significant interaction between valence and content, $F(1, 90) = 161.78$, $p < .001$. Post-hoc tests revealed that positive Warmth-related traits were rated more favourably ($M = 7.66$) than were positive Competence-related traits ($M = 6.65$), $F(1, 48) = 54.75$, $p < .001$, and that negative Warmth-related traits were rated more negatively ($M = 2.02$) than were negative Competence-related traits ($M = 3.93$), $F(1, 42) = 101.86$, $p < .001$. Overall, these results suggest that the Warmth dimension is more polarized in terms of favourability ratings. This finding is consistent with the behavioural adaptive approach (Peeters & Czapinski, 1990) and replicates previous research which suggested that the Warmth-dimension is more saturated with affect than the Competence dimension (Wojciszke et al., 1998b), due to its greater relevance for approach-avoidance.

Instances Confirming. The ANOVA yielded a significant main effect of valence, $F(1, 90) = 161.76$, $p < .001$, such that, consistent with previous findings (Rothbart & Park, 1986), positive traits required more instances of confirming behaviours ($M = 5.54$, $SD = .87$) to be established than did negative traits ($M = 3.36$, $SD = .92$). There was also a significant main effect of trait content, $F(1, 90) = 5.12$, $p = .026$; overall, Warmth-related traits required fewer instances of confirming behaviours ($M = 4.32$, $SD = 1.71$) to be established than did

Competence-related traits (M = 4.77, SD =.87). These results were qualified by a significant interaction between valence and content, $F(1, 90) = 27.65$, $p < .001$. Positive Warmth-related traits required more instances to be confirmed (M = 5.78) than did positive Competence-related traits (M = 5.29), $F(1, 48) = 4.17$, $p = .047$, and negative Warmth-related traits required fewer instances to be confirmed (M = 2.87) than did negative Competence-related traits (M = 4.08), $F(1, 42) = 32.54$, $p < .001$. Thus, the difference between positive and negative traits on the instances confirming dimension was more pronounced for Warmth than for Competence. However, positive traits required more instances to be confirmed than negative traits within each content category, $F(1, 50) = 143.85$, $p < .001$ and $F(1,40) = 37.73$, $p < .001$, respectively. These findings indicate a negativity effect for both Warmth and Competence, which was, however, less pronounced in the Competence domain. These results are not consistent with a positivity effect for Competence (i.e., more instances required to confirm negative than positive traits).

Because differences between positive and negative traits in global favourability were more pronounced for Warmth than for Competence, and because favourability is a major predictor of instances ratings (Rothbart & Park, 1986), the valence x content interaction could be due to differences in favourability. To investigate this possibility, we repeated the analysis adding trait favourability as a covariate. When favourability was controlled for, the valence x content interaction ceased to be significant, $F(1, 89) = 2.08$, $p = .152$. Positive Warmth and positive Competence did not differ in terms of instances confirming (adjusted M = 4.96 and 4.80, respectively), $F(1, 89) <1$, but the difference between negative Warmth and negative Competence remained significant (adjusted M = 3.87 and 4.48), $F(1,89) = 4.98$, $p = .031$.

Instances Disconfirming. We obtained a significant main effect of valence, $F(1, 90) = 68.16$, $p < .001$, with positive traits requiring fewer instances to be disconfirmed (M = 4.47, SD =.98) than negative traits (M = 5.83, SD =.79), in line with the notion that negative traits

are harder to 'lose' than positive traits (Rothbart & Park, 1986). This main effect was qualified by a significant interaction between valence and content, $F(1, 90) = 40.19$, $p < .001$. Simple effects analyses revealed that positive Warmth-related traits required fewer instances to be disconfirmed ($M = 3.87$) than did positive Competence-related traits ($M = 5.12$), $F(1, 48) = 34.08$, $p < .001$, and that negative Warmth-related traits required more instances to be disconfirmed ($M = 6.11$) than did negative Competence-related traits ($M = 5.41$), $F(1, 42) = 10.22$, $p = .003$. Moreover, while the difference between positive and negative traits on the instances disconfirming dimension was significant for Warmth-related traits, $F(1, 50) = 110.03$, $p < .001$, it was not significant for Competence-related traits, $F(1, 40) = 1.87$, $p = .173$, indicating a negativity effect for Warmth and the absence of a negativity effect (but no positivity effect) for Competence.

We also repeated the analysis with favourability partialled out. The significant valence x content interaction re-emerged in the analysis of covariance (ANCOVA), $F(1, 89) = 12.66$, $p = .001$. Simple effects tests controlling for trait favourability revealed a significant main effect of content for positive traits, $F(1, 47) = 15.18$, $p < .001$, such that positive Warmth-related traits required fewer instances to be disconfirmed (adjusted $M = 3.97$) than did positive Competence-related traits (adjusted $M = 5.17$). There was no significant difference between Warmth (adjusted $M = 5.99$) and Competence (adjusted $M = 5.37$) for negative traits when differences in favourability were controlled for.

Imaginability, Occasions, and Population Frequency. The same 2 x 2 ANOVA was run on the additional dimensions in order to explore differences between Warmth and Competence. As we had no specific hypotheses about Warmth-Competence differences on these dimensions, we adjusted the accepted significance level to $p < .01$. Only one significant result emerged: There was a significant main effect of valence for the population frequency dimension, $F(1, 90) = 32.26$, $p < .001$. In line with previous findings (Rothbart & Park,

1986), the occurrence of positive traits in the general population was rated as more frequent (M = 5.49, SD = .69) than the occurrence of negative traits (M = 4.65, SD = .76).

Intercorrelations of Dimensions

Next, we examined the intercorrelations among the dimensions, which are shown in Table 3. Because many of the variables share a substantial amount of variance with overall trait favourability, which is likely to result in spurious intercorrelations, we present both zero-order correlations as well as correlations with favourability partialled out.

Intercorrelations across all traits. Overall, the obtained correlations among the dimensions were, with a few exceptions, in line with the main findings reported by Rothbart and Park (1986). Consistent with the negativity bias, favourability was highly positively correlated with the instances confirming dimension (r = .84) and highly negatively correlated with the instances disconfirming dimension (r = -.69). Trait favourability was also highly correlated with its judged frequency in the population (r = .55), which is consistent with the Pollyanna Principle (Matlin & Stang, 1978) and the high frequency of positive events in general (Kanouse & Hanson, 1972). Moreover, consistent with previous results was the asymmetrical relation between the instances confirming and instances disconfirming dimensions (r = -.69), which remained significant when favourability was partialled out (r = -.30), and the symmetrical relationship for both the imaginability (r = .47) and occasions (r = .62) dimensions, which themselves were significantly correlated with each other (from r = .41 to r = .54). We examined one additional dimension in the current study, the word frequency in British English. This variable was significantly correlated with imaginability confirming and marginally significantly correlated with imaginability disconfirming (r = .29 and r = .18), suggesting that behavioural referents for trait words that are frequently used are more easy to imagine.

Differences between Warmth and Competence. Correlations between favourability and the instances dimensions calculated separately for Warmth and Competence are shown in Table 4. We tested whether correlations differed significantly between the two content categories using Fisher's *Z*-test for bivariate correlations (performed online; Preacher, 2005). The correlation between favourability and instances confirming was positive for both Warmth (r = .88) and Competence (r = .65), but significantly lower for Competence, z = 2.80, p = .005. The correlation between favourability and instances disconfirming was negative for both Warmth (r = -.84) and Competence (r = -.12), but significantly lower, z = 5.13, p < .001, and insignificant, for Competence. We also examined whether the asymmetrical relationship between instances confirming and instances disconfirming applies equally to Warmth and Competence. There was a strong negative correlation between instances confirming and instances disconfirming for Warmth (r = -.84) and a significantly reduced and insignificant negative correlation for Competence (r = -.17), z = -4.89, p < .001. The difference in correlations is similar when favourability is partialled out (partial r = -.42 for Warmth, and partial r = -.12 for Competence) and the difference between partial correlations was significant, $t(89)$ = 3.38, p = .001, when tested using a standard moderation test (Baron & Kenny, 1986).

Overall, the results of Study 2 indicate that the negativity bias applies primarily to the ascription of Warmth. For Competence, a reduced (for instances confirming) or absent (for instances disconfirming) negativity effect emerged. Thus, we can conclude that trait content does matter. The fact that no positivity effect was obtained for Competence could be due to a number of differences in methodology between the present research and previous research on the role of behaviour diagnosticity in trait ascription. Whereas Rothbart and Park's (1986) method assesses participant's beliefs about how they typically judge traits, in the studies conducted by Skowronski and Carlston (1987, 1992) and Reeder and colleagues (e.g.,

Reeder, 1979; Reeder & Spores, 1983) participants were presented with actual behaviours and the impact of those behaviours on actual trait ratings was assessed. Rothbart and Park's method quite possibly leads people to think about moderate behaviours that they typically encounter in everyday life (and that are therefore salient in memory); the positivity effect, however, is most evident when extreme behavioural information is presented (see Skowronski & Carlston, 1987).

Furthermore, Reeder and Brewer (1979) suggested that, within a specific situational context, the structure of trait-behaviour implications may be altered, such that the strength of some implicational links is enhanced and that of others reduced or eliminated. For example, when there are situational demands for competent behaviour, as is often the case, perceivers expect that most persons (regardless of their level of competence) will try to perform competently, so behaviours demonstrating both high and low competence may be seen as informative about a target's competence (see Reeder, Henderson, & Sullivan, 1982). In such a situation, relatively few instances of behaviour would be required to confirm incompetence, and a positivity effect is likely to emerge only for very high levels of competence. To explore this issue further, we assessed the likelihood of trait-inconsistent behaviour in Study 3.

The effect of personal involvement on trait judgments may further help to explain the present findings. Participants in the present study were probably thinking about common situations in their life involving people they usually interact with and whose actions could affect them. Skowronski and Carlston (1992) showed that, as personal involvement in a trait ascription task increases, trait judgments are less influenced by behaviour diagnosticity, and the positivity bias in judgments of intelligence is reduced. According to Neuberg and Fiske (1987), the presence of motivational goals could lead to enhanced weighting of negative information in impression formation. In Study 3 we explore whether trait attributes related to

motivational factors, specifically self-protection from harm and self-enhancement, could lead to bias in the confirmation of negative traits and the disconfirmation of positive traits.

Study 3: Aspects of Behaviour Diagnosticity, the Potential Harmfulness of Incorrect Trait Ascriptions, and Self-Desirability as Predictors of Trait (Dis-)confirmability

In this study we scaled our trait sample on five additional dimensions in order to further illuminate differences between Warmth and Competence and to explore the extent to which a number of different factors, related to the behavioural range of the trait possessor and the motivational goals of the perceiver, may underlie trait (dis-) confirmability.

First, we examined whether traits related to Warmth and Competence differ in terms of perceived behavioural range. To do this, we scaled traits on the perceived likelihood of trait-inconsistent behaviours, which is related to the diagnosticity of trait-related behaviours. As trait-behaviour implications are likely to be asymmetric in the domains of Warmth and Competence (Reeder & Brewer, 1979), we predicted that inconsistent behaviours would be judged to be more likely for positive Competence than for negative Competence, and more likely for negative Warmth-related traits compared to positive Warmth-related traits. Furthermore, since behaviours that are frequently and almost exclusively associated with members of one trait category are highly diagnostic (Skowronski & Carlston, 1987), this dimension should be related to the instances disconfirming dimension, such that more behavioural instances should be required to disconfirm a trait when inconsistent behaviours are common for that trait. This dimension is not directly related to the instances confirming dimension, but due to the asymmetry of the instances dimension (Rothbart & Park, 1986) and the hierarchical structure of trait-behaviour implications in the Warmth and Competence domains (Reeder & Brewer, 1979), a negative correlation can be expected.

We also tested whether different factors predict the perceived behavioural range of traits related to Warmth and Competence. Reeder (1993) proposed that whereas social

desirability issues as described by Jones and Davis (1965) are probably at least partly responsible for the asymmetrical diagnosticity of positive and negative behaviours in the Warmth domain, perceptions of performance limitations underlie the inference of Competence. People of low ability lack the power to bring about a high-level of performance whereas people of high ability can exhibit both low- and high ability performances. These differences in power, or controllability, should lead to the asymmetric trait-behaviour implications described by Reeder and Brewer's (1979) hierarchically restrictive schema. In this study we measured both the difficulty of pretending to have a trait as well as the difficulty of hiding a trait in order to assess the controllability aspect. We expected an interaction of trait valence with trait content for both these dimensions, such that positive and negative traits differ on these dimensions for Competence, but not for Warmth. It should be easy to pretend that one has negative Competence-related traits (e.g., it is easy to pretend that one is *stupid* or *clumsy*) but difficult to hide, whereas the opposite should apply to positive-Competence related traits. Moreover, we tested Reeder's proposition that controllability attributes affect behaviour diagnosticity ratings (which we operationalized as the likelihood of trait-inconsistent behaviour) only in the Competence domain.

Reeder (1993) also proposed that adaptive concerns (see also Peeters & Czapinski, 1990) could further underlie judgments of a person's Warmth. People form impressions of other people with the goal of predicting others' future behaviour and guiding their own behaviour towards others (e.g., avoiding or approaching them). Thus, a motive to protect oneself from people who lie, cheat, or behave aggressively could lead people to be conservative in trait ascriptions. Although it is not possible to address motivational factors directly with Rothbart and Park's (1986) methodology, scaling traits on attributes that may be relevant for motives can give some insight into the roles that they may play. To do this, we asked participants to judge traits according to how harmful they thought it could be to them

personally if they ascribed the given trait incorrectly to someone. We expected that incorrect trait ascriptions would be more harmful for positive than for negative traits. Generally, assigning a 'bad' person to a 'good' category (and then approaching them) carries greater risk and should be avoided more persistently than assigning a 'good' person to a 'bad' category (and then avoiding them). For example, erroneously believing that someone is trustworthy and then disclosing confidential information to them could have serious consequences, whereas erroneously believing that someone is *un*trustworthy and therefore *not* disclosing confidential information could not. Trait ascription could therefore be seen as a form of risk-taking, where costs typically have a greater deterrence value than gains have an attraction value (Kahnemann & Tversky, 1984). If the potential harmfulness of an incorrect trait ascription influences trait ascriptions, then this dimension should be correlated positively with the instances confirming, and negatively with the instances disconfirming dimension.

Moreover, we suggest that incorrect trait ascription could be harmful in both the Warmth- and the Competence domains, as both types of trait ascriptions could guide our behaviour towards others. However, in line with Peeters and Czapinski's (1990) ideas, we expected that an erroneous inference of positive Warmth-related traits has greater potential for harm than an erroneous inference of positive Competence-related traits. Other people's Warmth has a direct effect on the observer, whereas other people's Competence is more ambiguously related to consequences for the perceiver and depends more on special circumstances (Peeters & Czapinski, 1990; Wojciszke et al., 1998b).

The present study also sought to investigate the possibility that self-serving biases play a role in trait (dis-)confirmability. Festinger (1954) proposed that people have a drive to evaluate themselves on important dimensions and often do so by comparing themselves to relevant others. Subsequent research revealed that this comparison process is anything but unbiased; it is subject to a number of motives, of which self-enhancement, that is attaining or

maintaining positive self-esteem, is the strongest (Sedikides, 1993). People use a variety of strategies in order to arrive at favourable views of themselves. These include self-serving attributions about their own behaviour, but also biased judgments of others which then *in*directly enhance their own self-images (see Dunning, 2001, for a review). We believe that one possible strategy to ensure favourable social comparisons is to apply very stringent criteria for inferring that other people possess desirable traits (i.e., others have to engage in many trait-confirming behaviours to be ascribed a desirable trait and few behaviours to be ascribed an undesirable trait; and to engage in few trait-disconfirming behaviours before the perceiver decides that they do not possess a desirable trait and many trait-disconfirming behaviours before the perceiver decides that they do not possess an undesirable trait).

If this strategy serves self-enhancement motives, it should be especially pronounced for traits that are important to one's self-concept and self-esteem. Given that self-esteem is an indicator of the extent to which someone meets cultural standards (Tesser, 2001) it is likely that both one's Warmth and one's Competence are important for one's self-esteem. Warmth is important because being able to maintain good relationships with others ensures social approval and inclusion, and satisfies the need to belong (Baumeister & Leary, 1995). Competence, on the other hand, is important to self-evaluation because it is directly related to the ability to achieve goals and succeed in life (Peeters, 1983; Wojciszke et al., 1998b).

However, in a direct test of the relative importance of Warmth and Competence for the self-concept, Wojciszke et al. (1998b) revealed that Competence dominates the self-concept, is seen as more desirable for the self than Warmth, and has a greater impact on people's self-esteem than does Warmth (see Phalet & Poppe, 1997, for analogous findings for in-group vs. out-group stereotypes). Thus, the self-enhancement motive may be particularly important when judging others' Competence. In this study, we scaled traits on the extent to which they are seen as desirable for the self. We expected positive traits to be generally more

desirable than negative traits and, in line with Wojciszke et al. (1998b), that this difference would be more pronounced for Competence than for Warmth. Moreover, if self-enhancement motives play a role in the inference of other people's traits, we expected self-desirability ratings of traits to correlate positively with instances confirming ratings and negatively with instances disconfirming ratings.

Method

Participants

Participants were 53 (9 male, 42 female, 2 did not specify their gender; mean age = 22 years) students at a British university. Each dimension was rated by 10 to12 participants.

Procedure and Measures

The procedure was essentially the same as that of Study 2. Participants rated how difficult it is for someone to pretend to have the given trait if they do not actually possess it (*Difficulty Pretending*), how difficult it is for someone to pretend *not* to have the given trait if they actually do (*Difficulty Hiding*), how likely it is that someone who possesses the given trait shows trait-inconsistent behaviours (*Diagnosticity*), how harmful it could be to incorrectly ascribe the given trait (*Potential Harm*), or how desirable it would be to score high on this trait compared to other people (*Self-desirability*). Each participant rated all traits on *one* dimension only.

Results and Discussion

Reliability

The analyses closely resembled those of Study 2. Three unreliable judges were excluded. The numbers of original and final judges for each dimension along with coefficient

alpha values before and after deletion of unreliable judges appear in Table 5. The reliabilities were generally satisfactory, with final alphas ranging from .62 to .98.

Differences between Warmth and Competence on the Dimensions

Using all reliable judges, we computed average scores for each trait on each of the five dimensions. Table 6 presents mean ratings and standard deviations for positive and negative Warmth and positive and negative Competence. Results of a series of 2 (valence: positive vs. negative) x 2 (content: Warmth vs. Competence) ANOVAs on mean ratings for each dimension are reported below.

Difficulty Pretending. There was a significant main effect of valence, $F(1, 90) = 39.12$, $p < .001$, such that pretending to have positive traits was rated as being more difficult ($M = 4.86$, $SD = 1.15$) than pretending to have negative traits ($M = 3.75$, $SD = .97$). This effect was qualified by a significant interaction between valence and content, $F(1, 90) = 30.17$, $p < .001$. While participants rated pretending to have positive Competence-related traits as more difficult ($M = 5.59$) than pretending to have positive Warmth-related traits ($M = 4.18$), $F(1, 48) = 29.31$, $p < .001$, the opposite pattern emerged for negative traits: pretending to have negative Warmth-related traits was rated as being more difficult ($M = 4.03$) than pretending to have negative Competence-related traits ($M = 3.33$), $F(1, 42) = 6.16$, $p = .017$. As predicted, the difference between positive and negative traits was only reliable among Competence-related traits, $F(1, 40) = 62.94$, $p < .001$.

Difficulty Hiding. The ANOVA yielded a main effect of valence, $F(1, 90) = 108.43$, $p < .001$. Negative traits were rated as more difficult to hide ($M = 6.13$, $SD = 1.09$) than were positive traits ($M = 4.39$, $SD = .88$). This result was further qualified by a significant interaction between valence and content, $F(1, 90) = 27.75$, $p < .001$. Hiding negative Competence-related traits was rated as more difficult ($M = 6.88$) than hiding negative Warmth-related traits ($M = 5.60$), $F(1, 42) = 21.99$, $p < .001$. In addition, positive Warmth-

related traits were rated as being more difficult to hide (M = 4.68) than were positive Competence-related traits (M = 4.07), $F(1, 48) = 6.71$, $p = .013$. The difference between positive and negative traits was, as predicted, significant for Competence-related traits, $F(1, 40) = 146.32$, $p < .001$. It was also, unexpectedly, significant for Warmth, $F(1, 50) = 12.51$, $p = .001$. Overall, these findings indicate that, as expected, positive and negative Competence-related traits differed in terms of controllability, such that it is difficult to pretend to be competent but easy to hide, while it is easy to pretend to be incompetent, but difficult to hide.

Diagnosticity. The ANOVA yielded a significant main effect of trait content, $F(1, 90) = 9.54$, $p = .003$: Trait-inconsistent behaviours were rated as more likely for Competence-related traits (M = 4.86, SD = .75) than for Warmth-related traits (M = 4.34, SD = .74). However, this effect was qualified by a significant interaction between valence and content, $F(1, 90) = 15.08$, $p < .001$. Post-hoc tests revealed that trait-inconsistent behaviours were rated as significantly more likely for positive Competence-related traits (M = 5.16) than for positive Warmth-related traits (M = 3.91), $F(1, 48) = 41.98$, $p < .001$. However, the difference between negative Competence-related traits (M = 4.46) and negative Warmth-related traits (M = 4.76) only approached statistical significance, $F(1, 42) = 2.78$, $p = .103$. Moreover, consistent with Reeder and Brewer's (1979) hierarchically restrictive schema, inconsistent behaviours were rated significantly more likely for negative than for positive Warmth-related traits, $F(1, 50) = 25.86$, $p < .001$, whereas the opposite was found for Competence-related traits, $F(1, 40) = 10.79$, $p = .002$.

Potential Harm. We obtained main effects of valence, $F(1,90) = 243.36$, $p < .001$, and trait content, $F(1,90) = 49.54$, $p < .001$, as well as a significant interaction between valence and content, $F(1,90) = 37.64$, $p < .001$. Because the potential harm dimension was highly correlated with favourability, we repeated the analysis with trait favourability as a covariate. The obtained effects remained significant in the ANCOVA, $F(1, 89) = 14.50$, $p < .001$, $F(1,$

89) = 42.36, $p < .001$, and $F(1, 89) = 12.84$, $p < .001$, respectively. As expected, incorrectly ascribing positive traits was rated as more harmful ($M = 6.46$, $SD = 1.24$) than incorrectly ascribing negative traits ($M = 4.20$, $SD = .53$). The difference between positive and negative traits was significant for both Warmth, $F(1, 50) = 353.30.80$, $p < .001$, and Competence, $F(1, 40) = 30.76$, $p < .001$. Moreover, incorrectly ascribing Warmth-related traits was rated as more harmful ($M = 5.81$, $SD = 1.68$) than incorrectly ascribing Competence-related traits ($M = 4.90$, $SD = 1.03$). However, this difference was only significant for positive traits: incorrectly ascribing positive Warmth-related traits was rated as more harmful ($M = 7.37$) than incorrectly ascribing positive Competence-related traits ($M = 5.48$), $F(1, 48) = 69.44$, $p < .001$. This difference remained significant when differences in trait favourability were controlled for, $F(1, 47) = 21.10$, $p < .001$. There was no significant difference between negative Warmth-related ($M = 4.25$) and negative Competence-related traits ($M = 4.12$), $F(1, 42) < 1$, on this dimension. Overall, these results confirm the hypotheses that the incorrect ascription of positive traits can potentially be more harmful than the incorrect ascription of negative traits, and that this difference is more pronounced for Warmth than for Competence, consistent with Peeters and Czapinski's (1990) behavioural-adaptive model.

Self-desirability. The ANOVA yielded a significant main effect of valence, $F(1, 90) = 1660.36$, $p < .001$, such that, unsurprisingly, scoring high on positive traits compared to others was rated as more desirable ($M = 7.15$, $SD = .55$) than scoring high on negative traits ($M = 1.97$, $SD = .70$). This main effect was qualified by a significant interaction between valence and content, $F(1, 90) = 8.27$, $p = .005$. Positive Warmth-related traits were rated as more desirable ($M = 7.34$) than positive Competence-related traits ($M = 6.95$), $F(1, 48) = 6.57$, $p = .014$, but the difference between Warmth and Competence in the rated desirability of negative traits ($M = 1.83$ and $M = 2.17$, respectively) only approached statistical significance, $F(1, 42) = 2.70$, $p = .108$. When overall trait favourability was controlled for,

the valence x content interaction remained significant, $F(1, 89) = 26.45$, $p < .001$. However, Warmth-Competence differences *reversed*. Post-hoc tests controlling for favourability revealed that positive Competence-related traits were more, although only marginally significantly, desirable for oneself (adjusted $M = 7.32$) than positive Warmth-related traits (adjusted $M = 7.00$), $F(1, 47) = 3.49$, $p = .068$, and that negative Competence-related traits were less desirable for oneself (adjusted $M = 1.22$) than negative Warmth-related traits (adjusted $M = 2.48$), $F(1, 41) = 24.11$, $p < .001$. These findings support the hypothesis that Competence is more important for the self than is Warmth (Wojciszke et al., 1998b).

Intercorrelations of Dimensions

Next, we examined intercorrelations of these dimensions as well as their correlations with the instances dimensions assessed in Study 2. They are presented in Table 7. Since some of the dimensions were highly correlated with trait favourability, we also present partial correlations. Table 8 presents correlations calculated separately for Warmth and Competence. Not all significant correlations are discussed here. Rather, we will address a number of specific questions. First, we examined whether controllability aspects are related to the perceived likelihood of inconsistent behaviours (diagnosticity) for Competence, but not for Warmth (see Reeder, 1993). Overall, the controllability dimensions were negatively correlated with each other ($r = -.47$). Traits that were difficult to pretend to have were also easy to hide and vice versa. This correlation, however, was only significant for Competence ($r = -.72$) and significantly reduced for Warmth ($r = -.05$), $z = -4.00$, $p < .001$. As expected, and consistent with Reeder (1993), controllability aspects were only related to diagnosticity in the Competence domain. Traits that are difficult to pretend to have were also more likely to be associated with frequent inconsistent behaviours ($r = .34$). Moreover, the results suggest that traits that are difficult to hide are less likely to be associated with inconsistent behaviours ($r = -.52$). There were no significant correlations between the controllability dimensions and

the likelihood of inconsistent behaviours for Warmth and the differences between Warmth and Competence were significant, $z = 2.40$, $p = .016$ and $z = -3.39$, $p = .001$, respectively.

The likelihood of trait-inconsistent behaviours was significantly correlated with the instances dimensions. As expected, the more likely trait-inconsistent behaviours, the more inconsistent behavioural instances were required to disconfirm the trait ($r = .46$). However, this correlation differed significantly between Warmth and Competence, $z = 4.94$, $p < .001$. Whereas the likelihood of inconsistent behaviours was strongly and significantly correlated with instances disconfirming for Warmth ($r = .73$), there was no significant correlation between these two dimensions for Competence ($r = -.13$). These results did not change when favourability was partialled out. The likelihood of inconsistent behaviours was also significantly negatively correlated with instances confirming ($r = -.31$), suggesting that for traits for which inconsistent behaviour occurs often, fewer confirming behavioural instances are required to instantiate the trait. This relationship does not follow logically, but was expected due to the asymmetrical relationship between the instances dimensions and the hierarchically-structured nature of trait-behaviour implications in both the Warmth and Competence domains. This correlation was moderated by content, $z = -6.17$, $p < .001$: Only for Warmth did this dimension correlate negatively with instances confirming ($r = -.74$). There was a positive correlation between the likelihood of inconsistent behaviours and instances confirming for Competence ($r = .36$), which, however, disappeared when favourability was partialled out (partial $r = -.08$).

The potential harm of incorrect trait ascriptions was highly correlated with a number of dimensions due to its high correlation with overall trait favourability. As can be seen in Table 8, even after the overall trait favourability was partialled out, the potential harm of incorrect trait ascriptions was positively correlated with instances confirming (partial $r = .24$) and negatively with instances disconfirming (partial $r = -.49$), suggesting that the more

harmful the potential consequences of an incorrect ascription are, the more conservatively the trait is ascribed and the more easily it is disconfirmed. These correlations were moderated by content. The correlation between potential harm and instances confirming was significant for Warmth (partial $r = .49$), but not for Competence (partial $r = .28$), $t(89) = -2.04$, $p = .045$, consistent with Peeters and Czapinski's (1990) suggestion that the motive to avoid harm may play a greater role in ascriptions of Warmth. However, the correlation between potential harm and instances disconfirming was significant for both Warmth (partial $r = -.34$) and Competence (partial $r = -.48$), and stronger for Competence, $t(89) = 2.57$, $p = .011$.

A trait's desirability for the self was also highly correlated with many dimensions due to its high correlation with favourability. Thus, we will again discuss only partial correlations. There was no significant partial correlation of self-desirability with instances confirming overall ($r = -.07$), however, a significant positive correlation emerged for Competence (partial $r = .32$), but not for Warmth (partial $r = -.10$). This result suggests that the more desirable a trait is for oneself, the more instances are required to confirm the trait for other people, but only if the trait pertains to Competence. Nonetheless, the difference between Warmth and Competence was not significant, $t(89) < 1$. There was also no significant overall correlation between self-desirability and instances disconfirming. The correlations between self-desirability and instances disconfirming differed significantly for Warmth and Competence, $t(89) = 2.87$, $p = .005$, such that the negative correlation was stronger for Competence (partial $r = -.26$) than for Warmth (partial $r = -.02$), but neither was significant. These results provide only limited evidence for the prediction that self-serving biases may play a greater role for the ascription of Competence than Warmth. Overall, however, the present findings are consistent with the idea that motivational goals can lead to the enhanced weighing of negative information (Neuberg & Fiske, 1987). As both the self-protection from harm and self-enhancement dimensions were correlated with the instances dimensions, but uncorrelated

with each other when favourability was partialled out, we can conclude that both motives may bias trait ascriptions.

General Discussion

This research extended Rothbart and Park's (1986) classic work on the (dis-) confirmability of trait concepts by addressing the moderating role of trait content (Warmth vs. Competence), and by scaling traits on additional dimensions to gain further insights into the different factors underlying trait (dis-) confirmability. We will now discuss the theoretical contributions of this work, specifically relating to the role of trait content and the determinants of trait (dis-) confirmability, and the practical implications of these findings.

Theoretical Contributions and Directions for Future Research

Does Content Matter?

The results obtained in Study 2 suggest that content does matter. The strong negativity bias reported by Rothbart and Park (1986) was replicated for traits related to Warmth, but was greatly reduced for traits related to Competence, a finding that was only in part due to differences in overall favourability between the two content categories. Whereas positive traits in the Warmth and Competence domains were rated to require about equal amounts of confirming evidence to be established, negative traits in the Warmth domain seem to be inferred more easily than negative traits in the Competence domain. Moreover, positive Competence-related traits were judged to require more trait-inconsistent behavioural instances to be disconfirmed than positive Warmth-related traits, and negative Competence-related traits were judged to require fewer instances to be disconfirmed than negative Warmth-related traits. Also, the positive correlation between trait favourability and the instances confirming dimension and the negative correlation between trait favourability and instances disconfirming, which where highly significant for Warmth, were significantly

reduced among Competence-related traits. Although these results imply that trait favourability is a less important predictor of trait (dis-) confirmability for Competence than for Warmth, they are not consistent with a positivity effect. Rather, the findings of Study 2 are consistent with a reduced negativity effect or null effect in the Competence domain. As discussed earlier, these findings may be due to a number of methodological differences between the present study and prior research on trait ascriptions (e.g., Skowronski & Carlston, 1987).

However, further insights into the effects of trait content on trait (dis-) confirmability may be gained when the present results are examined from the perspective of the Five-Factor Model (FFM) of personality. Using a lexical approach to identify the domains that are most important in describing the self and others, a series of studies have emphasized Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Intellect/Openness as five broad domains of personality (see Goldberg, 1993). An inspection of the present trait list using a classification of traits according to the FFM (Goldberg, 1982) revealed that our Warmth category broadly overlaps with Agreeableness, and that most of our Competence-related traits relate to Conscientiousness, and some to Intellect/Openness. We examined these three types of trait content in terms of their (dis-)confirmability in an exploratory analysis. The nature of the analyses goes beyond the scope of the current paper; however, it should be noted that a strong negativity effect emerged for Agreeableness ($N = 42$, $r = .88$, $p < .001$, for instances confirming, and $r = -.90$, $p < .001$, for instances disconfirming), a reduced negativity effect emerged for Conscientiousness ($N = 25$, $r = .70$, $p < .001$, for instances confirming, and $r = -.64$, $p = .007$, for instances disconfirming), and there was a null effect ($N = 9$, $r = .34$, n.s., for instances confirming) and a marginally significant *positivity* effect ($r = -.64$, $p = .064$, for instances disconfirming) for the Intellect/Openness category. These findings are only exploratory and should be replicated in future research using a larger number of

representative traits for each content category. However, these results strongly suggest that future research on the effect of trait content on trait (dis-) confirmability may benefit from further breaking down the Competence category into Conscientiousness and Intellect/Openness, which seem to have different patterns of trait (dis-) confirmability.

Future work could also investigate the (dis-) confirmability of Extraversion and Emotional Instability. According to Reeder (1993), Extraversion has an ability component and it is thus possible that a positivity effect would emerge (i.e., extraverted behaviours are seen as informative and lead to correspondent inferences). However, because most social situations demand extraverted behaviour, introverted behaviour might be seen as highly informative and therefore lead to a negativity bias (Jones & Davis, 1965). For Neuroticism, which is the domain that is most closely associated with social desirability (see Funder & Dobroth, 1987), a negativity bias is likely to emerge. Identifying biases in the ascription of these different types of traits could have important implications for the diagnosis of personality disorders.

Furthermore, Coker, Samuel, and Widiger (2002) showed that the high poles of these five personality domains contain some socially undesirable and probably maladaptive traits (e.g., '*ingratiating'* at the high pole of Agreeableness and '*overbookish'* at the high pole of Conscientiousness) and that low poles contain some socially desirable traits. This suggests that Warmth and Competence may not be perfectly correlated with social desirability and that both content and social desirability could be systematically varied in future studies on trait ascription. This would, for example, allow researchers to disentangle the independent effects of behaviour diagnosticity (e.g., Skowronski & Carlston, 1987) and social desirability (Jones & Davis, 1965) at the trait level.

Determinants of Trait (Dis-) Confirmability

Rothbart and Park (1986) examined personality traits along relevant dimensions which determine a trait's potential (dis-)confirmability; these were related to the observability of trait-related behaviour, the link between observed behaviour and inferred disposition, and the structure of the social environment. One factor, operationalized as the imaginability of trait-related behaviour, represents the extent to which trait-related behaviour is observable and likely to be recognized as such. It is determined by specific characteristics of a trait, such as its abstractness or level of generality, and whether it implies clear behavioural referents or whether behavioural exemplars are accessible in memory. In the present research this factor was associated with the frequency with which trait words are used in language. The second factor associated with characteristics of the trait is that of how confirming and disconfirming behaviours (once observed) are used to make trait judgements, i.e., the dimensions assessing the number of behavioural instances required to confirm or disconfirm a trait. Although this factor is likely to be, in part, determined by the schematic representation of trait-behaviour implications (see Reeder & Brewer, 1979), it is also dependent on the characteristics of the social environment and the goals of the perceiver.

The frequency of a trait in the general population and the frequency of social occasions allowing trait-related behaviours represent the environmental factor; this factor is related to the possibility of observing trait-related behaviour in everyday social situations. One major determinant of this factor is that of social desirability: socially desirable behaviour is performed frequently (because it is rewarded) and socially undesirable behaviour is performed infrequently (because it is punished). This has implications for trait ascriptions. Behaviour which violates social norms is particularly informative about some underlying stable quality of the actor and will thus lead to dispositional attributions, while socially desirable behaviour, which is likely to be performed irrespective of whether an actor

possesses the corresponding trait, is less informative (Jones & Davis, 1965). Consistent with these ideas, trait favourability emerged as a strong predictor of instances confirming and disconfirming in both Rothbart and Park's (1986) and the present research.

The present work further extended Rothbart and Park's (1986) analysis by showing that how behavioural information is used is not just determined by the social desirability of a trait, but also by the perceived behavioural range of the trait-possessor and the motivational goals of the perceiver. Three dimensions – the difficulty in pretending and hiding a trait and the likelihood of trait-inconsistent behaviours – assessed the behavioural range of the trait-possessor (see Reeder & Brewer, 1979) and are related to the diagnosticity of trait-related behaviour (Skowronski & Carlston, 1987). In line with Reeder and Brewer (1979), the likelihood of trait-inconsistent behaviour was perceived to be greater for traits related to negative Warmth compared to positive Warmth, and for traits related to positive Competence compared to negative Competence. Consistent with Reeder's (1993) ideas, this indicator of behavioural range was predicted by the controllability of confirming and disconfirming behaviours, i.e., the extent to which a trait can be hidden or 'faked', in the Competence but not in the Warmth domain. It was also correlated with instances confirming and disconfirming, suggesting that the perceived behavioural range of the trait possessor has some impact on trait ascriptions.

There was also evidence from the present research that motivational goals of the perceiver, specifically self-protection from harm and self-enhancement, may lead to increased weighting of negative behavioural information in trait ascription. In line with Peeters and Czapinski's (1990) behavioural adaptive approach, the potential harm dimension correlated positively with the number of behavioural instances required to confirm a trait and negatively with the number of behavioural instances required to disconfirm a trait. This indicates that the harm that can result from incorrect trait ascriptions may bias people's decision criteria for

ascribing traits to others (see also Kahnemann & Tversky, 1984). Also, a trait's desirability for the self was found to correlate positively with instances confirming and negatively with instances disconfirming, but only among Competence-related traits. These results suggest that self-serving motivations may bias trait inferences in the Competence but not the Warmth domain. In sum, the current analysis suggests that the ease with which a trait is (dis-) confirmed is determined by at least three types of variables: specific characteristics of the trait that characterize its link to discrete concepts, conditions that characterize the social environment, and characteristics of the social perceiver making the inferences. This analysis could be further extended in future work by investigating variables that characterize the target individual about whom inferences are to be made, such as the target's group membership, status, and relationship to the perceiver, all of which are likely to moderate the effects of the dimensions explored here and elsewhere. Thus, a complete model of trait (dis-)confirmability needs to delineate the factors that are most appropriate for describing the links between trait, environment, target, and perceiver.

Nonetheless, some of the present findings need to be further consolidated by experimental research examining trait inferences under more naturalistic circumstances for two main reasons. First, the current results are only correlational and can therefore not imply causality. Second, Rothbart and Park's (1986) methodology assesses general *beliefs* about attributes of traits, not actual trait attributions on the basis of observed behaviour (but see Funder & Dobroth, 1987, on the significant relationship between these trait attributes and accuracy in personality judgment). This methodology also does not permit controlling for behaviour extremity, an important factor that moderates the extent to which negativity and positivity effects are shown (Skowronski & Carlston, 1987), or to address motivational factors more directly. We thus suggest that future research simultaneously manipulates factors that affect behaviour diagnosticity (e.g., situational demands, behaviour extremity)

and motivational forces (e.g., personal involvement, threat manipulations to increase the motivation to avoid potentially costly trait ascriptions) under more naturalistic circumstances. This would allow researchers to assess the unique contributions of both the diagnosticity (Reeder & Brewer, 1979; Skowronski & Carlston, 1987) and behavioural-adaptive (Peeters & Czapinski, 1990) perspectives on trait attribution.

Implications of the Present Findings

An interesting implication of these findings concerns the perspective of the target individuals to whom these traits are ascribed. That negative traits are assigned relatively easily, are lost or revised with some difficulty, are difficult to hide, and are assigned with little expectation that such an assignment will produce negative consequences, helps to explain why so much energy is invested in managing the impressions that other people have of us. The long term risks of being assigned a negative trait, or set of traits, are potentially very high. Research on the negative effects of stigma (e.g., Crocker, Voelkl, Testa, & Major, 1991) and of prejudice from the target's perspective (e.g., Swim, Scott, Sechrist, Campbell, & Stangor, 2003) readily attests to this. Furthermore, these risks are likely to be exacerbated for members of cohesive social networks, or high entitativity groups, wherein behavioural norms are more strongly enforced.

For interventions aimed at improving negative perceptions of out-groups, such as intergroup contact, the present model and findings imply that contact is likely to have more of an effect on certain kinds of stereotypical traits and beliefs, and less of an effect on others. Further compounding this analysis is the finding that high and low prejudiced individuals employ different attributional processes in behaviour-trait inferences, resulting in greater stereotype resistance among the highly prejudiced (see Sherman, Stroessner, Conrey, & Azam, 2005). The present findings also suggest that that the (dis-)confirmability of traits may be further biased by the presence of motivational goals such as avoiding harm and obtaining

positive self-views, both of which may also be influential in intergroup situations where protecting the in-group from harmful outsiders (see Brewer, 2001) as well as attaining positive distinctiveness in comparison to other groups (e.g., Tajfel & Turner, 1979) are important goals. Future attempts to understand the trait-inference process in intergroup relations should therefore incorporate both the content of the stereotypes as well as the nature of the perceiver (viz., prejudice level, motivational goals) into the model.

Conclusions

The present research extends previous work on the (dis-)confirmability of trait concepts by examining the moderating effect of trait content (Warmth vs. Competence), and by scaling traits on a number of attributes that provide insights into the different processes that may underlie trait ascriptions. The present results indicate that Warmth and Competence show different patterns of (dis-) confirmability. This research also provided preliminary evidence that a number of different factors, related to attributes of the trait, the environment, and the perceiver, may simultaneously determine trait ascriptions. By considering the various determinants of trait (dis-)confirmability in chorus, and by distinguishing different types of trait content, we may arrive at a more complete understanding of the trait inference process.

References

Anderson, N. H. (1968). Likableness ratings of 555 personality-trait words. *Journal of Personality and Social Psychology*, *15*, 79-103.

Baumeister, R. F., Bratslavsky, E., Finkenhauer, C., Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*, 323-370.

Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, *117*, 497-529.

Brewer, M. B. (2001). Ingroup identification and intergroup conflict: When does ingroup

love become outgroup hate? In R. Ashmore, L. Jussim, & D. Wilder (Eds.), *Social identity, intergroup conflict, and conflict reduction*. New York: Oxford University Press.

Coker, L.A., Samuel, D.B., & Widiger, T.A. (2002). Maladaptive personality functioning within the Big Five and the Five-factor Model. *Journal of Personality Disorders, 16*, 385-401.

Crocker, J., Voelkl, K., Testa, M., & Major, B. (1991). Social stigma: The affective consequences of attributional ambiguity. *Journal of Personality and Social Psychology*, *60*, 218-228.

Dunning, D. (2001). On the motives underlying social cognition. In A. Tesser & N. Schwarz (Eds.), *Blackwell handbook of social psychology: Intraindividual processes* (pp. 348-374). London: Blackwell.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, *7*, 117-140.

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, *82*, 878-902.

Funder, D. C., & Dobroth, K. M. (1987). Differences between traits: Associated with interjudge agreement. *Journal of Personality and Social Psychology*, *52*, 409-418.

Gilbert, D. T. (1998). Ordinary personology. In D.T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *Handbook of social psychology* (4th ed., pp. 89-150). Boston: McGraw Hill.

Goldberg, L.R. (1982). From ace to zombie: Some explorations in the language of personality. In C.D. Spielberger & J.N. Butcher (Eds.), *Advances in personality assessment* (Vol. 1, pp. 203-234) Hillsdale, NJ: Erlbaum.

Goldberg, L.R. (1993). The structure of phenotypic personality traits. *American Psychologist, 48*, 26-34.

Jones, E. E., & Davis, K. E. (1965). From acts to dispositions: The attribution process in person perception. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 2, pp. 219-266). New York: Academic Press.

Kahnemann, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist*, *39*, 341-350.

Kanouse, D. E., & Hanson, L. R. (1972). Negativity in evaluations. In E. E. Jones, D. E. Kanouse, H. H. Kelley, R. E. Nisbett, S. Valins, & B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior* (pp. 47-62). Morristown, NJ: General Learning.

Kelk, B. (2003). UK English wordlist. *Available at: http://www.bckelk.uklinux.net*.

Matlin, M. W., & Stang, D. J. (1978). *The pollyanna principle.* Cambridge, MA: Schenkman.

Neuberg, S. L., Fiske, S. T. (1987). Motivational influences on impression formation: Outcome dependency, accuracy-driven attention, and individuating processes. *Journal of Personality and Social Psychology*, *53*, 431-444.

Peeters, G. (1983). Relational and informational patterns in social cognition. In W. Doise & S. Moscovici, S. (Eds.), *Current issues in European social psychology* (Vol. 1, pp. 201-237). Cambridge: Cambridge University Press.

Peeters, G., & Czapinski, J. (1990). Positive-negative asymmetry in evaluations: The distinction between affective and informational effects. In W. Stroebe & M. Hewstone (Eds.), *European review of social psychology* (Vol. 1, pp. 33-60). New York: Wiley.

Phalet, K., & Poppe, E. (1997). Competence and morality dimensions of national and ethnic stereotypes: a study in six eastern-European countries. *European Journal of Social Psychology*, *27*, 703-723.

Pratto, F., & John, O. (1991). Automatic vigilance: The attention-grabbing power of negative social information. *Journal of Personality and Social Psychology*, *61*, 380-391.

Preacher, K. J. (2005). Calculation for the test of the difference between two independent

correlation coefficients. *Available at: www.unc.edu/~preacher/corrtest/corrtest.htm*.

Reeder, G. D. (1979). Context effects for attributions of ability. *Personality and Social Psychology Bulletin*, *5*, 65-68.

Reeder, G. D. (1993). Trait-behavior relations and dispositional inference. *Personality and Social Psychology Bulletin*, *19*, 586-593.

Reeder, G. D., & Brewer, M. B. (1979). A schematic model of dispositional attribution in interpersonal perception. *Psychological Review*, *86*, 61-79.

Reeder, G.D., Henderson, D.J., & Sullivan, J.J. (1982). From dispositions to behaviours: The flip side of attribution. *Journal of Research into Personality, 16*, 355-375.

Reeder, G. D., & Spores, J. M. (1983). The attribution of morality. *Journal of Personality and Social Psychology*, *44*, 763-745.

Rosenberg, S., Nelson, C., & Vivekananthan, P. S. (1968). A multidimensional approach to the structure of personality impressions. *Journal of Personality and Social Psychology*, *9*, 283-294.

Rosenberg, S., & Sedlack, A. (1972). Structural representations of implicit personality theory. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 6, pp. 235–297), New York: Academic Press.

Rothbart, M., & Park, B. (1986). On the confirmability and disconfirmability of trait concepts. *Journal of Personality and Social Psychology*, *50*, 131-142.

Sedikides, C. (1993). Assessment, enhancement, and verification determinants of the self-evaluation process. *Journal of Personality and Social Psychology*, *65*, 317-338.

Skowronski, J. J., & Carlston, D. E. (1987). Social judgment and social memory: The role of cue diagnosticity in negativity, positivity, and extremity biases. *Journal of Personality and Social Psychology*, *52*, 689-699.

Skowronski, J. J., & Carlston, D. E. (1992). Caught in the act: When impressions based on

highly diagnostic behavious are resistant to contradiction. *European Journal of Social Psychology*, *22*, 435-452.

Sherman, J., Stroessner, S., Conrey, F., & Azam, O. (2005). Prejudice and stereotype maintenance processes: Attention, attribution, and individuation. *Journal of Personality and Social Psychology*, *89*, 607-622.

Swim, J., Scott, E., Sechrist, G., Campbell, B., & Stangor, C. (2003). The role of intent and harm in judgments of prejudice and discrimination. *Journal of Personality and Social Psychology*, *84*, 944-959.

Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In W. G. Austin & S. Worchel (Eds.), *The psychology of intergroup relations* (pp. 33-48). Monterey, CA: Brooks/Cole.

Tesser, A. (2001). Self esteem. In A. Tesser & N. Schwarz (Eds.), *Blackwell handbook of social psychology: Intraindividual processes* (pp. 479-498). London: Blackwell.

Wojciszke, B. (2005). Morality and competence in person and self perception. *European Review of Social Psychology, 16*, 155-188.

Wojciszke, B., Bazinska, R., & Jaworski, M. (1998a). On the dominance of moral categories in impression formation. *Personality and Social Psychology Bulletin*, *24*, 1245-1257.

Wojciszke, B., Dowhyluk, M., & Jaworski, M. (1998b). Moral and competence-related traits: how do they differ? *Polish Psychological Bulletin*, *29*, 283-294.

Author Note

This research was funded by a doctoral studentship from the University of Oxford awarded to the first author. We would like to thank Michael Ashton and Lewis Goldberg for providing us with factor loadings and trait ratings of the Five-factor Model trait list. We would also like to thank Glenn Reeder and an anonymous reviewer for comments on an earlier draft of this paper. Correspondence concerning this paper should be addressed to Nicole Tausch at the Department of Experimental Psychology, University of Oxford, South Parks Road, Oxford, OX1 3UD, UK. E-mail: nicole.tausch@psy.ox.ac.uk

Table 1. Number of Judges and Coefficient Alpha Values Before and After Deletion of Unreliable Judges (Study 2)

Dimension	Original *N*	Original α value	Final *N*	Final α value
Favourability	10	.98	10	.98
Instances Confirming	10	.87	10	.87
Instances Disconfirming	11	.80	10	.84
Imaginability Confirming	10	.69	10	.69
Imaginability Disconfirming	10	.66	10	.66
Occasions Confirming	10	.83	10	.83
Occasions Disconfirming	9	.65	8	.68
Population Frequency	10	.74	10	.74

Table 2. Means and Standard Deviations of Trait Ratings on 8 Dimensions as a Function of Trait Content and Valence (Study 2)

	Trait Content			
	Warmth		Competence	
Dimension	Positive	Negative	Positive	Negative
Favourability	7.66 (.58)	2.02 (.49)	6.65 (.35)	3.93 (.67)
Instances Confirming	5.78 (1.00)	2.87 (.73)	5.29 (.62)	4.08 (.65)
Instances Disconfirming	3.87 (.84)	6.11 (.70)	5.12 (.66)	5.41 (.74)
Imaginability Confirming	5.72 (1.16)	6.13 (1.25)	6.10 (.92)	5.85 (1.00)
Imaginability Disconfirming	6.26 (.98)	6.01 (.94)	6.12 (.85)	6.61 (.68)
Occasions Confirming	5.89 (1.20)	5.75 (1.22)	5.58 (.91)	5.64 (.80)
Occasions Disconfirming	5.50 (1.02)	5.79 (.89)	5.31 (.70)	5.49 (.58)
Population Frequency	5.69 (.67)	4.71 (.76)	5.28 (.66)	4.57 (.77)

Table 3. Zero-order Intercorrelations of Dimensions and Correlations with Favourability Partialled Out across Traits ($N = 94$) (Study 2)

Dimensions	Instances Confirming	Instances Disconfirming	Imagine Confirming	Imagine Disconfirming	Occasions Confirming	Occasions Disconfirming	Population Frequency	Word Frequency
Favourability	.84***	-.69***	-.05	.05	.06	-.15	.55***	.07
Instances Confirming	-	-.69***	-.28*	-.08	-.11	-.27**	.42***	.02
Partial favourability	-	*-.30***	*-.45****	*-.23**	*-.30***	*-.26**	*-.10*	*-.07*
Instances Disconfirming		-	.26*	-.12	-.05	.03	-.34**	-.05
Partial favourability			*.31***	*-.11*	*-.01*	*-.11*	*.06*	*.01*
Imaginability Confirming			-	.47***	.44***	.41***	.17	.29**
Partial favourability				*.48****	*.44****	*.41****	*.23**	*.30***
Imaginability Disconfirming				-	.53***	.54***	.20^{+}	.19^{+}
Partial favourability					*.53****	*.55****	*.20^{+}*	*.18^{+}*
Occasions Confirming					-	.62***	.52***	.20^{+}
Partial favourability						*.64****	*.58****	*.20^{+}*
Occasions Disconfirming						-	.15	.17
Partial favourability							*.29***	*.18^{+}*
Population Frequency							-	.19^{+}
Partial favourability								*.18^{+}*

Note. $^{+}$ $p < .10$; * $p < .05$; ** $p < .01$; *** $p < .001$.

Table 4. Intercorrelations of Favourability, Instances Confirming, and Instances Disconfirming and Partial Correlations, calculated separately for Warmth (N = 52) and Competence (N = 42) (Study 2)

Dimension	Instances Confirming	Instances Disconfirming
Favourability		
Warmth	.88***	-.84***
Competence	.65***	-.12
Instances Confirming		
Warmth	-	-.84***
Competence	-	-.17
Instances Confirming (*partial favourability*)		
Warmth	-	*-.42***
Competence	-	*-.12*

Note. ** $p < .01$; *** $p < .001$.

Table 5. Number of Judges and Coefficient Alpha Values Before and After Deletion of Unreliable Judges (Study 3)

Dimension	Original *N*	Original α value	Final *N*	Final α value
Difficulty Pretending	10	.77	10	.77
Difficulty Hiding	10	.71	9	.79
Diagnosticity	12	.61	11	.62
Potential Harm	11	.83	10	.89
Self-desirability	10	.98	10	.98

Table 6. Means and Standard Deviations of Trait Ratings on Additional Dimensions as a Function of Trait Content and Trait Valence (Study 3)

	Trait Content			
	Warmth		Competence	
Dimension	Positive	Negative	Positive	Negative
Difficulty Pretending	4.18 (.84)	4.03 (1.00)	5.59 (.99)	3.33 (.78)
Difficulty Hiding	4.68 (.89)	5.60 (.99)	4.07 (.76)	6.88 (.73)
Likelihood of Inconsistent Behaviour (Diagnosticity)	3.91 (.67)	4.76 (.63)	5.16 (.69)	4.46 (.66)
Potential Harm	7.37 (.75)	4.25 (.39)	5.48 (.85)	4.12 (.69)
Self-desirability	7.34 (.50)	1.83 (.79)	6.95 (.55)	2.17 (.49)

Table 7. Zero-order Intercorrelations of Additional Dimensions and Correlations with Favourability, Instances Confirming, and Instances Disconfirming, and Correlations with Favourability Partialled Out across Traits (N = 94) (Study 3)

Dimensions	Difficulty Hiding	Diagnosticity	Potential Harm	Self-desirability	Favourability	Instances Confirming	Instances Disconfirming
1. Difficulty Pretending	-.47***	.18^{+}	.20^{+}	.41***	.30**	.29**	.04
Partial favourability	*-.39****	*.25**	*-.05*	*.43****	-	*.07*	*.30***
2. Difficulty Hiding	-	-.19^{+}	-.45***	-.62***	-.48***	-.36***	.22*
Partial favourability		*-.33***	*-.13*	*-.61****	-	*.10*	*-.17*
3. Diagnosticity		-	-.40***	-.10	-.19^{+}	-.31**	.46***
Partial favourability			*-.41****	*.30***	-	*-.28***	*.46****
4. Potential Harm			-	.76***	.77***	.73***	-.75***
Partial favourability				*.14*	-	*.24**	*-.49****
5. Self-Desirability				-	.95***	.79***	-.63***
Partial favourability					-	*-.07*	*.13*

Note. $^{+}$ $p < .10$; * $p < .05$; ** $p < .01$; *** $p < .001$.

Table 8. Intercorrelations of Additional Dimension, Favourability, Instances Confirming, and Instances Disconfirming, and Correlations with Favourability partialled out, calculated separately for Warmth (N = 52) and Competence (N = 42) (Study 3)

	Zero-order Correlations							*Partial Correlations*					
Dimensions	Hide	Diagno	Harm	Self	Favour	InstCon	InstDis	Hide	Diagno	Harm	Self	InstCon	InstDis
Difficulty Pretending													
Warmth	-.05	-.16	.18	.02	.04	.22	-.02	-.04	-.17	.38**	-.13	.39**	.02
Competence	-.72***	.34*	.48**	.80***	.78***	.43**	-.03	-.15	-.09	.03	.24	-.16	.10
Difficulty Hiding													
Warmth	-	.15	-.42**	-.39**	-.39**	-.35*	.29*	-	-.12	-.15	.03	.00	-.08
Competence	-	-.52***	-.61***	-.91***	-.87***	-.56***	.20	-	-.22	-.22	-.54***	-.00	.19
Diagnosticity													
Warmth		-	-.66***	-.59***	-.58***	-.74***	.73***		-	-.42**	-.15	-.59***	.55***
Competence		-	.29^{+}	.54***	.50***	.36*	-.13		-	-.02	.27^{+}	.06	-.08
Potential Harm										-	.09	.49***	-.34*
Warmth			-	.92***	.92***	.90***	-.84***			-	.19	.28	-.48**
Competence			-	.62***	.60***	.56***	-.45**						
Self-Desirability													
Warmth				-	.99***	.86***	-.82***				-	-.10	-.02
Competence				-	.97***	.68***	-.18				-	.32*	-.26

Note. $^{+}$ $p < .10$; * $p < .05$; ** $p < .01$; *** $p < .001$; Hide = difficulty with which someone can hide having this trait; Diagn = likelihood of inconsistent behaviours (diagnosticity); Self = desirability of trait for oneself; Favour=Favourability; InstCo=Instances Confirming; InstDis=Instances Disconfirming.